Public Opinion Quarterly, Vol. 74, No. 3, Fall 2010, pp. 551–569

THE POLLS–REVIEW
INACCURATE AGE AND SEX DATA IN THE CENSUS PUMS FILES: EVIDENCE AND IMPLICATIONS

J. TRENT ALEXANDER
MICHAEL DAVERN
BETSEY STEVENSON*

Abstract We discover and document errors in public-use microdata samples ("PUMS files") of the 2000 Census, the 2003–2006 American Community Survey, and the 2004–2009 Current Population Survey. For women and men age 65 and older, age- and sex-specific population estimates generated from the PUMS files differ by as much as 15 percent from counts in published data tables. Moreover, an analysis of labor-force participation and marriage rates suggests the PUMS samples are not representative of the population at individual ages for those age 65 and over. PUMS files substantially underestimate labor-force participation of those near retirement age and overestimate labor-force participation rates of those at older ages. These problems were an unintentional byproduct of the misapplication of a newer generation of disclosure-avoidance proce-

J. TRENT ALEXANDER is a Research Scientist at the Minnesota Population Center at the University of Minnesota, Minneapolis, MN, USA. MICHAEL DAVERN is the Vice President and Director of Public Health Research at NORC at the University of Chicago, Chicago, IL, USA. BETSEY STEVENSON is an Assistant Professor of Business and Public Policy at the Wharton School at the University of Pennsylvania, Philadelphia, PA, USA. The authors would like to thank seminar participants at Wharton, the Minnesota State Demographers' Brownbag Series, and the University of Pennsylvania Population Studies Center; participants at the 2009 Joint Statistical Meetings and the 2010 Population Association of America annual conference; as well as Orley Ashenfelter, Carolyn Liebler, Peter Graven, David Johnson, Steven Ruggles, Stephen Tordella, and Justin Wolfers for useful discussions. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by NORC at the University of Chicago. This work was supported by the National Institutes of Health–National Institute for Child Health and Human Development [R01 HD43392, R01 HD043392-03S1, R24 HD041023 to J.T.A. and M.D.]; the National Institutes of Health–National Institute on Aging [P30 AG12836 to B.S.]; and the Boettner Center for Pensions and Retirement Security at the University of Pennsylvania and National Institutes of Health–National Institute of Child Health and Human Development Population Research Infrastructure Program at the University of Pennsylvania [R24 HD-044964 to B.S.]. *Address correspondence to Betsey Stevenson, 1454 Steinberg Hall–Dietrich Hall, 3620 Locust Walk, University of Pennsylvania, Philadelphia, PA 19104, USA; e-mail: betsey.stevenson@wharton.upenn.edu.

doi: 10.1093/poq/nfq033 Advance Access published on August 13, 2010

dures carried out on the data. The resulting errors in the public-use data could significantly impact studies of people age 65 and older, particularly analyses of variables that are expected to change by age.

Introduction

This article investigates serious problems with the age and sex data provided by the Census Bureau in many recent public data products. Census Bureau data resources include published tables based on the full data that the agency collects, and public-use microdata samples ("PUMS files") based on an anonymized subsample of the data that has been subjected to disclosure-avoidance techniques. For women and men age 65 and older, age- and sex-specific population estimates generated from many recent PUMS files differ substantially from counts presented in published data tables that were created using the full, confidential data. For example, population estimates from the 2000 decennial census PUMS files differ from the published data tables by up to 15 percent for some age and sex combinations. These differences are not related to regular sampling variation. Instead, the problems were an unintentional byproduct of the misapplication of disclosure-avoidance procedures carried out on the data. The resulting errors in the public-use data are severe, and, as such, we argue that the Census Bureau's PUMS files from several years should not be used to study people age 65 and older.

We discover and document the impact of these errors in PUMS files of the 2000 Census, the 2003–2006 American Community Survey (ACS), and the 2004–2009 Current Population Survey (CPS). We explore this issue with two main goals in mind. First, Census Bureau PUMS files are extremely important datasets for researchers and policymakers. The problematic data have been used by thousands of researchers for a variety of purposes. We aim to raise awareness of how the data errors could generate biases in current and future research results. Second, this issue provides an important cautionary tale for producers and users of data that are subject to disclosure-avoidance techniques. The newest generation of disclosure-avoidance techniques has significant benefits for both data producers and data users. Older techniques simply removed detail from datasets (by aggregating small categories, top-coding extreme values, etc.). Newer techniques, such as swapping or blanking, retain detail *and* provide better protection of respondents' confidentiality. However, the effects of the new techniques are less transparent to data users, and mistakes can easily be overlooked. Therefore, these new techniques carry increased responsibility for both data users and data producers to vigilantly review the anonymized data.

Our analysis begins with a review of available documentation on the issue, followed by a data analysis illustrating the apparent error in PUMS files from the decennial census, the American Community Survey, and the Current Population Survey. We then discuss the continuing effect that these problems are

likely to have on the study of aging and the elderly. We conclude by suggesting potential workarounds and longer-term approaches to correcting the errors.

Disclosure-Avoidance Techniques

PUMS files have always been subject to a wide range of disclosure-avoidance techniques. Some of these techniques are transparent to data users and are discussed at length in dataset documentation. The most common and well known of these are: (1) the microdata released to the public is only a sample of all the records the Census Bureau has; (2) the data file does not release low-level geographic identifiers; (3) variables with hundreds of categories have a smaller number of categories on the public-use file; and (4) continuous variables with outlying values, such as income and transportation time, are top- and bottom-coded.

Disclosure-avoidance techniques have grown more complex in the past few years, due largely to concerns that new data technologies present a growing disclosure threat. These newer techniques include swapping or rank swapping (also called switching), replacing randomly selected records with imputed values (as if the data were missing), and noise addition.[1]

Following communication with the authors of this paper, the Census Bureau acknowledged in two user notes that disclosure-avoidance techniques have caused seemingly minor problems with age and sex data in census and ACS PUMS files published between 2000 and 2007. The first user note was added in October 2008, in the errata notes that are appended to the Census 2000 PUMS codebook. The codebook's "Data Note 12" warns researchers that disclosure-avoidance techniques resulted in "some abnormal ratios for the number of men to the number of women (sex ratio) for people age 65 and over (U.S. Bureau of the Census 2008)." The note presents sex ratios and population counts by grouped years of age, comparing the 2000 one-percent PUMS file to published estimates from Summary File 3. Released in April 2009, ACS PUMS User Note 47 presents 2006 ACS data and contains identical warnings regarding the 2000–2006 ACS samples and the 2005–2007 ACS three-year sample (U.S. Bureau of the Census 2009). The Census 2000 user note indicates that "the PUMS files will not be rereleased using a modified technique as that would pose a disclosure risk."[2]

1. Details regarding disclosure-avoidance techniques used by the Census Bureau are discussed in Zayatz (2005) and Office of Management and Budget (2005).

2. On December 18, 2009, the Census Bureau released a revised version of the 2006 ACS data. Our analyses use the original version of the 2006 ACS data. The revised data were created by using a rank-swapping technique on the original ages of individuals age 65 and older. These new synthetic data have been analyzed by the Census Bureau, but the results are confidential. It is therefore too early to tell how successfully the synthetic data addresses the problems discussed here.

The user notes do not provide information on the particular technique that caused the problem, but the data itself can be used to better understand the effects that the errors could be having on research.[3] In the next section, we use public census data to describe the extent of the problem by comparing the faulty age and sex distribution of the PUMS files to those from published census tables.

Data Comparisons

The Census Bureau publishes tables that are almost always based on datasets that are much larger than the PUMS files. In the 2000 Census, published tables are based on either the complete population (in Summary Files 1 and 2) or about one-sixth of the population (in Summary Files 3 and 4). The largest PUMS file in 2000 included less than one-third of the cases that were used to make Summary Files 3 and 4. Similarly, American Community Survey PUMS files include about two-thirds of the cases in the larger datasets that are used to make the published tables available on the Census Bureau's American FactFinder site (http://factfinder.census.gov). While it is possible that the Census Bureau's internal files were also exposed to disclosure-avoidance techniques, these techniques may be done separately, to a lesser extent, or not at all.

2000 DECENNIAL CENSUS DATA

Documentation of the erroneous data focuses on problems with sex ratios in the PUMS files. However, the full extent of the problem can be seen more clearly in single-year population estimates by age and sex. In figure 1, we show sex- and age-specific population estimates from the 2000 five-percent Census PUMS as a proportion of published counts from Summary File 4. For all ages 0–64, PUMS estimates are never more than 1.5 percent higher or lower than published counts. In contrast, for those age 65 and up, PUMS estimates differ from published population counts substantially, in some cases by more than 10 percent. For instance, the PUMS estimate of the number of 65-year-old women is 895,052, which is only about 83 percent of Summary File 4's published count of 1,079,328.[4]

AMERICAN COMMUNITY SURVEY DATA

There are no published single-year age estimates from the American Community Survey (ACS), so it is not possible to view the problem at the same level

3. The Census Bureau has been unwilling to give further detail regarding the nature of the misapplication of the disclosure-avoidance techniques, since public knowledge of such detail may jeopardize the effectiveness of these techniques.

4. All PUMS data from Census 2000 and the ACS were based on microdata obtained from Ruggles et al. (2009), hereafter referred to as "IPUMS-USA." PUMS data from the CPS were based on microdata obtained from King et al. (2009), referred to in table citations as "IPUMS-CPS."

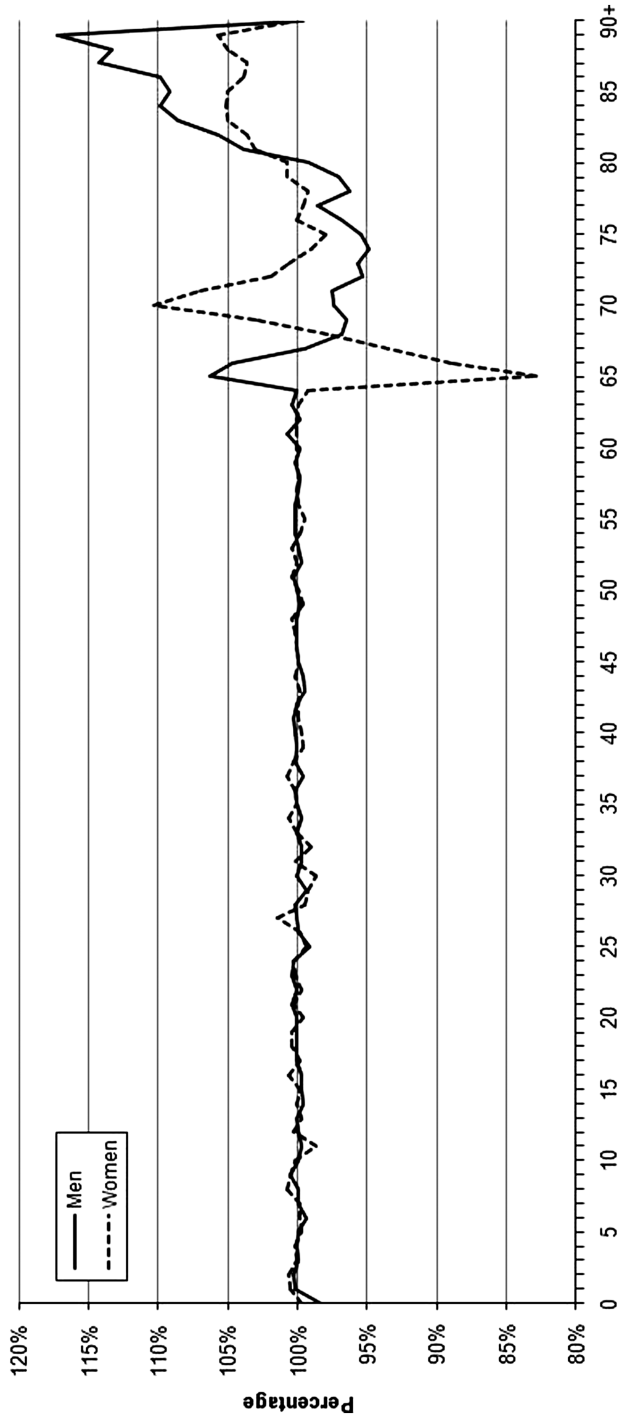

Figure 1. Population estimates from 2000 five-percent Census PUMS as a percentage of Census 2000 published data. Sources: Published population counts are from Census 2000 Summary File 4, Table PCT3 (http://factfinder.census.gov); population estimates are calculated using Census 2000 five-percent sample, IPUMS-USA (http://usa.ipums.org/usa/).

of detail as in the 2000 Census. The ACS publishes counts by age groups, and thus we compare these published group counts with population count estimates for comparable age groups calculated using population-weighted PUMS data. Once again we see that the PUMS calculations diverge substantially from published counts beginning at age 65. Figure 2 shows estimates from the 2006 ACS PUMS as a proportion of the 2006 ACS published data obtained from American FactFinder.

The problem with the disclosure-avoidance techniques was corrected before the release of the 2007 ACS PUMS file. Both the 2007 and 2008 ACS PUMS files produce estimates of the population at all ages that are close to those in the published tables. Examining the ACS PUMS since the inception of the ACS reveals that the problem with the disclosure-avoidance techniques in the ACS PUMS files appeared for the first time in 2003. PUMS estimates of the population of those age 65 years and older differ substantially from published accounts in the 2003–2006 survey years. The Census Bureau's Data Note 47 suggests that the errors also exist in the 2000–2002 ACS PUMS data; however, our analysis suggests that the 2000–2002 files do not contain errors.

Figure 3 shows ACS PUMS estimates as a proportion of published data for all ACS samples. To facilitate easy comparisons of the seven different samples, figure 3 presents data for women. The samples in figure 3's panel A—the 2001, 2002, 2007, and 2008 ACS—all produce good estimates of the female population at all ages. The samples included in panel B—the 2003–2006 ACS—reveal undercounts of women in their mid-60s and overcounts of women over age 85 in all samples. Additionally, in 2003 and 2004, the ACS PUMS samples overcount women in their early 70s and early 80s; in 2005 and 2006, ACS PUMS samples overcount women in their late 60s. The 2003 and 2004 ACS samples produce the worst estimates, particularly for 65- and 66-year-olds. In these samples, the PUMS estimates of 65- and 66-year-old women are about 85 percent of the count in the published tables. A similar analysis of the data for men reveals that estimates of the male population are also distorted in the 2003–2006 samples.

CURRENT POPULATION SURVEY DATA

While the Census Bureau has thus far acknowledged a link only between the misapplication of disclosure-avoidance techniques and data problems in the 2000 Census and the ACS public-use files, an investigation of the Current Population Survey (CPS) data suggests that the problems affecting estimates of the older population may have affected these public-use files as well.

There are no published population totals for the CPS, so it is not possible to conduct the same type of analysis we did for the 2000 Census and the

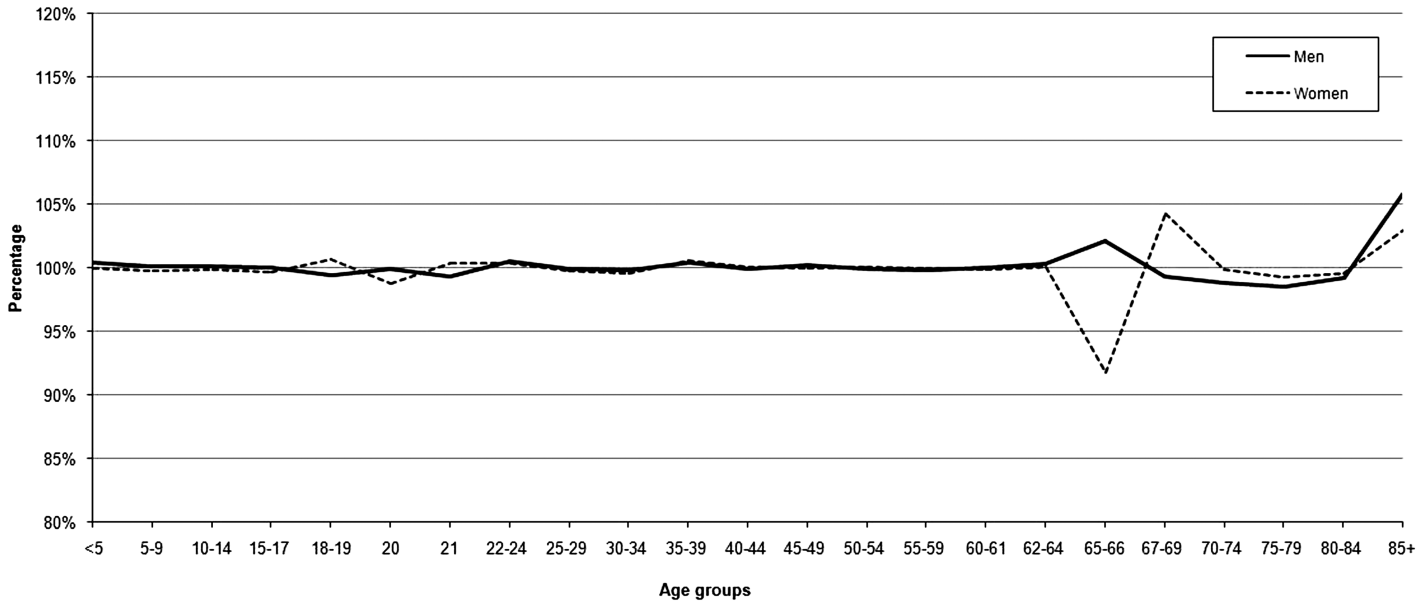

Figure 2. Population estimates from ACS 2006 PUMS as a percentage of ACS 2006 published data. Sources: Published data population counts are from 2006 ACS Table B01001 (http://factfinder.census.gov); population estimates are calculated using 2006 ACS PUMS, IPUMS-USA (http://usa.ipums.org/usa/).

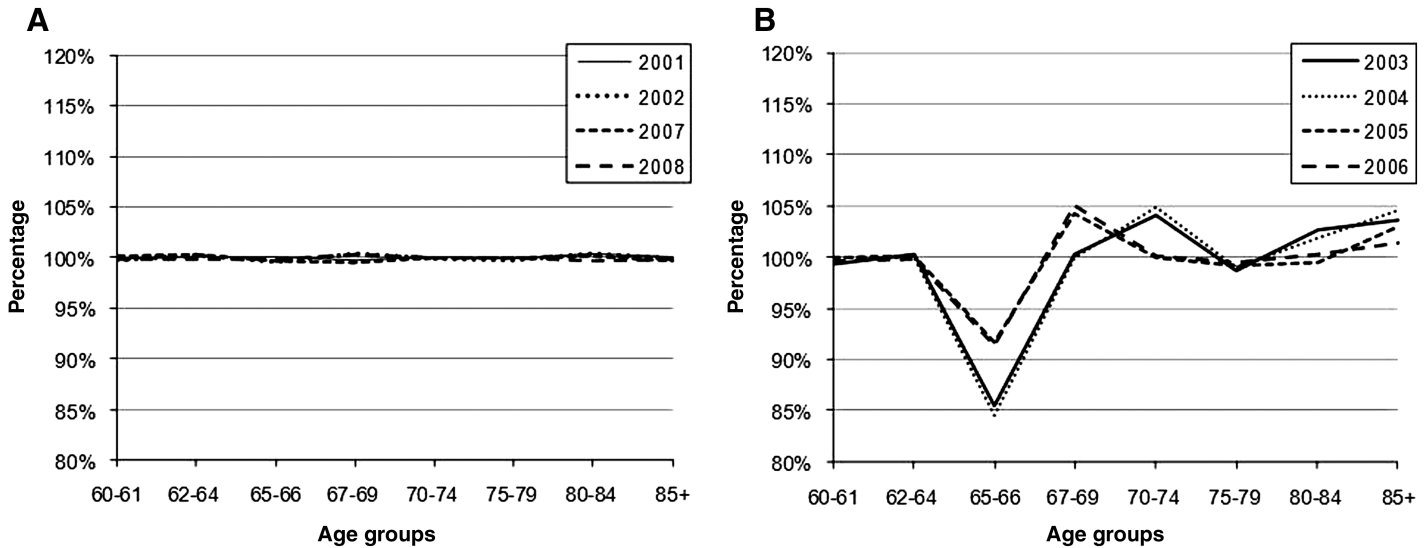

Figure 3. ACS PUMS estimates as a percentage of ACS PUMS published estimates, women only. Sources: Published population counts are from ACS Table B01001 (2004–2007); ACS Table P004 (2002–2003); 2001 Supplementary Survey Table P004 (http://factfinder.census.gov); ACS PUMS estimates are calculated using 2001–2008 ACS PUMS, IPUMS-USA (http://usa.ipums.org/usa/).

ACS.[5] Instead of comparing CPS estimates to an external standard, we investigate the problem in CPS using a measure that is internal to the dataset: the sex ratio. In the Census Bureau and ACS public-use files, the over- and undercounts of men and women age 65 and older do not occur proportionally by age and gender; indeed, at age 65 men tend to be overcounted while women tend to be undercounted. These under- and overcounts of men and women impact the sex ratio, resulting in implausibly large shifts in the sex ratio across the ages for those age 65 and up.

Panel A of figure 4 shows the sex ratio of men to women among 62-to-64-year-olds not living in group quarters, comparing three data sources: (1) the CPS Annual Social and Economic Supplement (CPS ASEC) PUMS; (2) the ACS PUMS; and (3) the ACS published data.[6] These data on 62-to-64-year-olds show the level of variation one would normally expect to see among the three different data sources, absent any disclosure-avoidance distortions. All three data sources produce similar estimates—among 62-to-64-year-olds there are about nine men for every 10 women. Not surprisingly, there is more variance in the estimates produced from the CPS, since the CPS is a much smaller sample than the ACS. Still, 90-percent confidence intervals around the CPS estimates easily contain estimates from the published ACS tables and ACS PUMS samples for each year. In other words, there are never statistically significant differences between the three sources' estimates of the 62-to-64-year-old population.[7]

In panel B of figure 4, we turn to those age 65 and 66, ages shown to have substantial errors stemming from the misapplication of disclosure-avoidance techniques in the ACS PUMS files from 2003 to 2006. As with the estimates for 62-to-64-year-olds, in 2001 and 2002, and again in 2007 and 2008, the ratio of men to women among 65- and 66-year-olds in the PUMS files tracks the ratios calculated from the published ACS data quite closely. Moreover, the ratios follow a sensible pattern given death patterns by age, with a small decline in the sex ratio at ages 62 to 64 from 0.9 to around 0.87 at ages 65 and 66.

5. ACS published tables are not directly compared to the CPS. The CPS excludes institutional group quarters, while the 2000–2004 ACS excludes all group quarters (institutional and non-institutional) and the 2005–2008 ACS includes all group quarters. There are also slight differences in how CPS and ACS identify non-institutional group quarters (ACS counts many more non-institutional group quarters than the CPS). Because of these differences, ACS published data never represent the same population that is sampled in the CPS. However, a comparison of population estimates by age still reveals unexpectedly large shifts in the ratio of CPS PUMS estimates to ACS published estimates at age 65 beginning in 2004.

6. Published data from the 2005–2009 ACS includes a small number of group-quarters cases; it was not possible to remove them.

7. We generated 90-percent confidence intervals for the CPS sex ratios using the delta method; standard errors were adjusted to take account of added weighting variance in light of the complex sample design of the CPS.

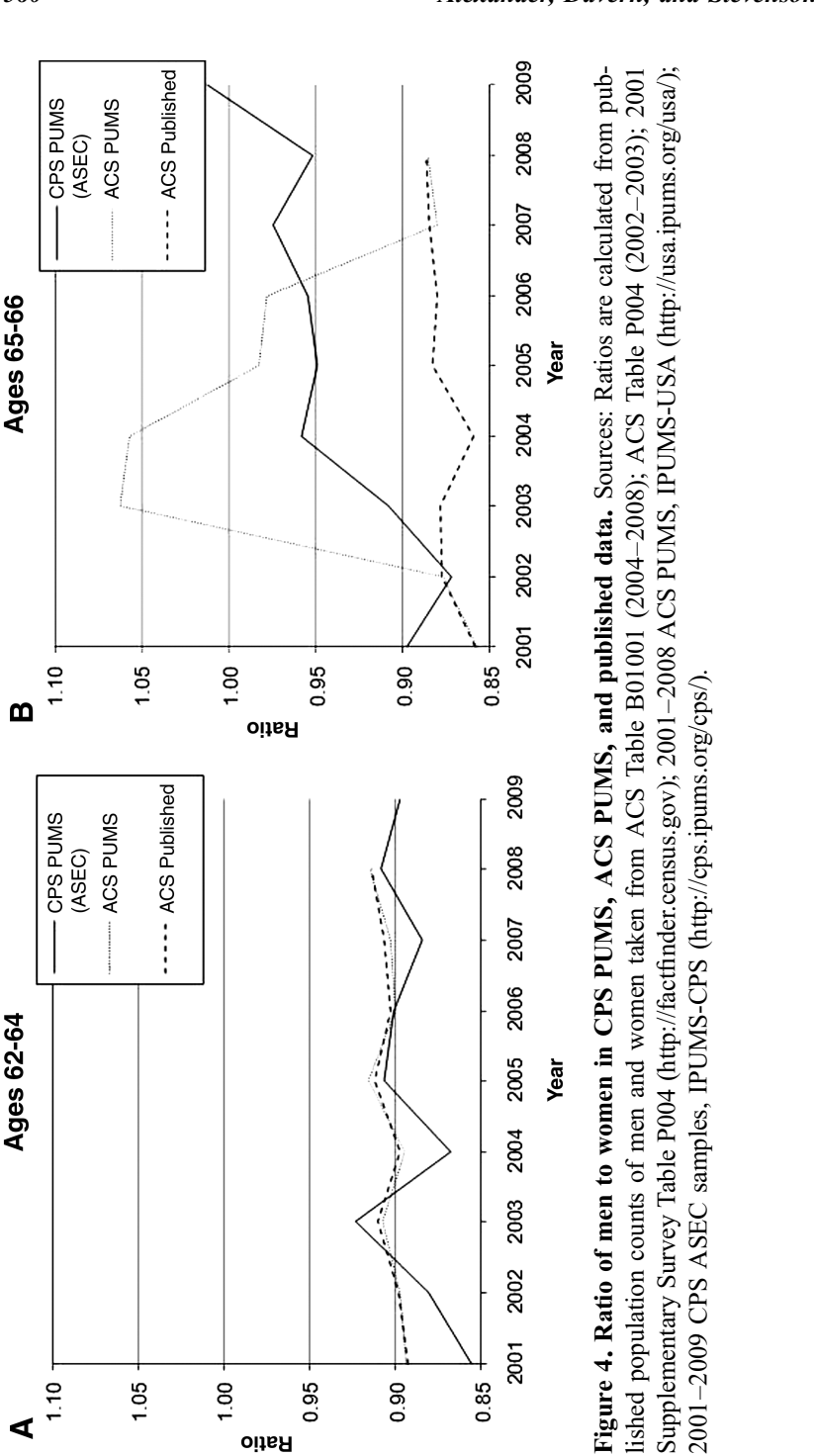

Figure 4. Ratio of men to women in CPS PUMS, ACS PUMS, and published data. Sources: Ratios are calculated from published population counts of men and women taken from ACS Table B01001 (2004–2008); ACS Table P004 (2002–2003); 2001 Supplementary Survey Table P004 (http://factfinder.census.gov); 2001–2008 ACS PUMS, IPUMS-USA (http://usa.ipums.org/usa/); 2001–2009 CPS ASEC samples, IPUMS-CPS (http://cps.ipums.org/cps/).

In contrast, the ratios calculated using the ACS PUMS data from 2003 to 2006 differ substantially from the ratios calculated using the published tables for this period. Moreover, these estimates differ from what one would expect given life expectancy patterns and the sex ratios for those a few years younger, with the gender ratio increasing, rather than decreasing, as these cohorts age. Indeed, in 2003 and 2004, the ACS PUMS estimates suggest a reverse gender ratio, with more men than women for those age 65 and 66.

Sex ratios from the CPS ASEC PUMS are also much higher than the published ACS data from 2004 to 2009. To test whether sex ratios in CPS data and published ACS data could be due to ordinary sampling error, we generated 90-percent confidence intervals for the CPS sex ratios. In 2000 through 2003, the published ACS sex ratios were contained well within the 90-percent confidence interval of the sex ratio estimated using the CPS PUMS data. Beginning in 2004, the 90-percent confidence interval no longer contains, in most years, the published ACS sex ratio. Moreover, it is unlikely that regular sampling variance would cause such large overestimates of the sex ratio of 65- and 66-year-olds calculated from the CPS data for six consecutive years. While there is no ACS data with which to compare the CPS in 2009, the 2009 CPS ASEC PUMS estimate of the sex ratio is the highest estimate seen across all years.

Taken together, the estimates in figure 4 suggest that the CPS ASEC PUMS samples may have been created using the Census Bureau's faulty disclosure-avoidance techniques beginning with the 2004 PUMS, continuing through the most recent available ASEC PUMS in 2009.

Thus far we have shown that the PUMS samples produce inaccurate population estimates and sex ratios for those age 65 and up. The larger issue with these erroneous data is not simply that sex ratios could be inaccurate, but rather that the age data attached to each case are probably often wrong, which creates the potential for the errors to spill over into analyses of any related variable.[8] When the PUMS vastly underestimates the number of 65-year-old women, for instance, we have to wonder whether these "missing" women are included elsewhere in the PUMS file. Given the unexplained surplus of women in higher age ranges, the altered cases were likely allocated to other, older ages. For instance, the lack of women in their mid-60s in figure 1 seems to be offset by spikes of women in their early 70s and 80s. However, it is difficult to know for sure what is driving the surplus at some ages and the deficit at others.[9]

8. It is possible that the problem is with the sex variable; however, the Census Bureau's corrected files did not alter the sex of any observation and did reassign ages for those age 65 and up. While the Census Bureau has not stated definitively that the problem resulted from rank swapping involving the age variable, it is most likely that this is what occurred.

9. The problems that we identify are not, however, driven by problems with the weights. Both weighted and unweighted data show similar problems. Moreover, the Census Bureau in private communication with the authors made it clear that the problem is not with the weights.

These "reassigned" age data could lead researchers to draw incorrect inferences about any variable related to age. For example, figure 5 shows PUMS estimates of labor-force participation in 2000 as a proportion of published rates from the Census Bureau's internal files. PUMS files underestimate labor-force participation of men and women age 65 and substantially overestimate labor-force participation of women in their 70s and older and of men over age 75. Thus, in the application of the disclosure-review techniques, people were accidentally assigned to groups whose actual labor-force participation is quite different from their own. For example, if the variable that was changed was age, then women that the PUMS identifies as 70-to-74-year-olds are actually women from an age group with higher labor-force participation, such as women in their 60s.

An examination of marriage rates further demonstrates problems with the population sample at individual ages and gender beginning at age 65. Figure 6 shows estimates of the proportion of women of each age who are currently married. In the 2007 ACS PUMS data, which does not have the identified errors, women's marriage rates decline steadily from age 60 to 75 as they become increasingly likely to be widows. In the 2006 ACS PUMS, which suffers from the misapplication of disclosure-avoidance techniques, marriage rates are particularly low for 65-year-old women and particularly high for 68-year-old women.

Figure 6's estimates for 65-year-old women reveal that the women whose ages were not reassigned (i.e., those who were still classified as 65-year-olds) had significantly lower marriage rates than one would expect based on the 2007 data. This illustrates that, beyond the fact that there are not enough women in the PUMS files at age 65, the women there are not representative of the whole set of 65-year-old women. This suggests that a *non-random* subset of 65-year-old women were reassigned to older ages. For example, most persons with an age value of 65 in the PUMS file may actually be a different age—there is no reason to believe that, just because there is an undercount, the remaining sample is truly the age or sex that they are assigned in the PUMS data. Clearly, the problems in Census Bureau and ACS age/sex data are complex, and analyses of any variable that is expected to grow or decline with each year of age could be biased as a result of the misapplied disclosure-avoidance techniques.

Who Is Affected by This Problem?

We suspect that the problem with age and sex data for older adults is not widely known among analysts relying on these data. After being alerted to the problem in 2008, the Census Bureau created data notes (Data Note 12 and User Note 47) to let users know of potential problems. However, these notes are not in locations where data users will necessarily see them. Moreover, the emphasis of the data notes could give the impression that this is an

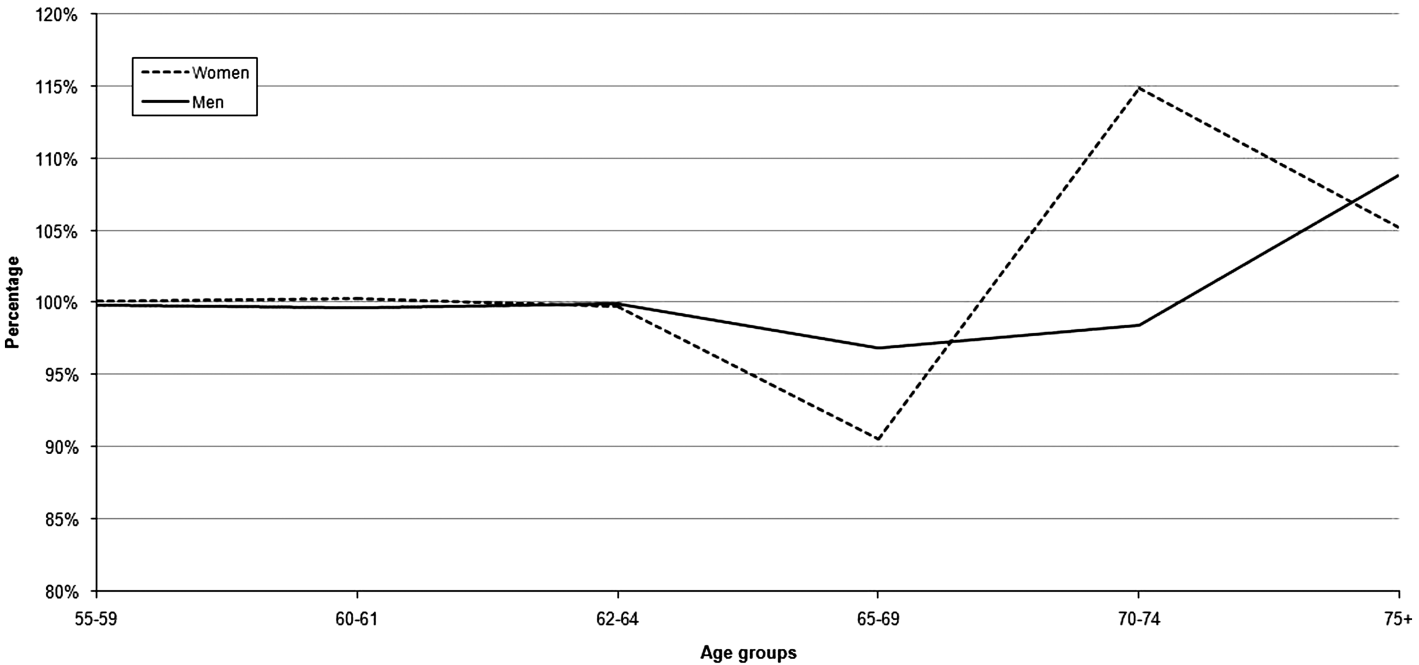

Figure 5. Labor-force participation estimates by age and sex, 2000 five-percent PUMS as a percentage of 2000 Census published data. Sources: Labor-force population counts are taken from Census 2000 Summary File 4, Table PCT70 (http://factfinder.census.gov); labor-force estimates are calculated using Census 2000 five-percent sample, IPUMS-USA (http://usa.ipums.org/usa/).

Figure 6. Percentage of women who were married in the 2006 and 2007 ACS. Sources: Percent of women married is calculated using the 2006 and 2007 ACS PUMS, IPUMS-USA (http://usa.ipums.org/usa/). Percent married includes only those married with a spouse present. Similar results are obtained for calculations that included spouses who are absent.

issue only for research that is specifically focused on sex ratios among the elderly.[10] Finally, users have spent nearly a decade using these data sources without being made aware of the problem. We believe that the problems could be having broad-ranging effects for several different groups of stakeholders, discussed in turn below.

Researchers who treat age as a continuous variable: These data issues raise serious concerns for researchers who analyze those age 65 and over in the PUMS data. The problems with the data go beyond that which is currently described in the Census Bureau note (which implies that only sex ratios will be erroneously estimated in PUMS files). As we document above, the issue is not simply that sex ratios could be inaccurate, but rather that the sex and age data attached to each case are probably often wrong, and the remaining sample is not likely representative of the actual population at a given age and sex. It is clear that PUMS files from the affected years should not be used by researchers for detailed studies of the 65-and-over population. Our research suggests that the data are accurate only when grouped into a single 65-and-over category. For example, our analysis of labor-force participation data shows that researchers interested in studying how the current economic downturn has affected retirement behavior will be led astray if they use ACS (2003–2006) or CPS PUMS (2004–2009) files in their analysis.

Social-service agencies that rely on PUMS data for important policy research: The second major group of stakeholders affected by faulty age and sex values is social-service agencies that rely on the PUMS. Census Bureau microdata products are widely used by the policy research community and by government agencies, both for allocation formulas to fund specific government programs and for more general program planning (Blewett and Davern 2007). One important example of estimates that could be impacted by the underlying error in the PUMS data are the simulations used to project the solvency of Old Age, Survivors, and Disability Insurance (OASDI). Age-sex ratios are commonly used to derive program estimates and for long-term forecasting for the OASDI Trustees Report.[11]

More generally, the Congressional Budget Office relies upon Census Bureau microdata to examine current programs and to assess the impact of programmatic changes associated with proposed legislation (Glied, Remler, and Zivin 2002). For those analyses of various policy options for populations

10. The Census Bureau's User Notes may focus on sex ratios because the authors of this paper stressed this issue when describing the problem to census staff. Our subsequent research to understand the extent of the problem revealed that sex ratios are one of many statistics that is impacted by the misapplication of disclosure-review techniques.
11. http://www.ssa.gov/OACT/TR/2009/V_economic.html#189335.

age 65 and over (e.g., Medicare and Social Security), it is possible that estimates using the affected census microdata could be adversely affected.

In many cases, public-use microdata from the decennial census and American Community Survey are the only sources that can address the needs of these agencies. While the most recent two ACS PUMS files do not have these errors, it is important to note that using public-use data for the 2000 Census or the 2003-to-2006 ACS may result in incorrect calculations and provide an incorrect baseline for understanding long-term change. For example, when a federal agency identifies significant change between the 2000 and 2010 Censuses, for instance, it may be interpreting social change where the only real change has been in the accuracy of the data.

Survey researchers who use PUMS data to generate population estimates: Federal government agencies, survey-data-collection vendors, and pollsters regularly use census data to construct survey weights. Tabulations generated from census PUMS data are often used in a process called "post-stratification," where preliminary survey weights are set so that the sample cases sum up to Census Bureau totals by geography, age, sex, race, ethnicity, and other key demographic variables (Office of Management and Budget 2001). Post-stratification is commonly implemented using a technique known as "raking," which could be sensitive to the age and sex errors in the PUMS. Raking allows survey researchers to fit many control variables to their preliminary weighted total to make sure the survey adds up to appropriate totals by, for example, age and sex within a specific geography. The data errors in the Census Bureau's PUMS data products would be problematic especially for those surveys that fit age by sex control totals from Census PUMS categories in five-year increments and include people 65 years of age and older.

Researchers using any weighted dataset in which the weighting strategy relied on the 2000 Census or the 2003–2006 ACS PUMS samples need to consider the potential for errors in their estimates. In such cases, researchers can investigate the potential error by charting any basic demographic pattern by single year of age or by small age groupings, looking for unexpected peaks or troughs in the statistics. As we have suggested above, the problems are severe enough that they tend to produce visible divergences from expected patterns, such as was seen in sex ratios and in sex-specific distributions of labor-force status or marital status. Alternatively, one could compare current estimates with those produced by reweighting using information from a census PUMS file that does not have errors (such as the 2001 ACS PUMS file).

Researchers, policymakers, and survey researchers have expressed a clear desire for accurate year-specific age data in the PUMS files. In fact, many researchers and policymakers spoke to this exact issue as the Census Bureau was preparing the 2000 PUMS data. At that time, the Census Bureau was considering a plan that would significantly reduce the PUMS's single-year age detail for persons over age 65. In the quotes below, just a small sample

of responses to a survey administered and published by the Census 2000 Advisory Committee (Census 2000 Advisory Committee 2000), researchers from a variety of backgrounds made their objections to this plan clear.[12]

- "Aggregated age especially at older ages would be disastrous for analyses of the older and oldest-old (one of the fastest-growing segments of the population)." (p. 35)
- "My research is on the marriage patterns of older women (especially related to the Social Security remarriage penalty). I need to calculate marriage rates by age for women ages 55–75, and this would be impossible with age-grouped data." (p. 51)
- "To understand age-related changes of behaviors and characteristics, the AGE IN SINGLE YEAR (65, 66, 67, etc.) is essential." (p. 58)
- "With the aging of the population the key demographic issue in the U.S., grouping data at the oldest ages poses a key threat to research on aging." (p. 61)

Data users need to be aware that, despite the Census Bureau's subsequent decision to provide year-specific age data for the elderly, the misapplication of disclosure-avoidance techniques has resulted in PUMS files that contain data that must be grouped into a single 65-and-over age group in order to provide accurate estimates of the population and behavior within it.

Correcting the Data

The most straightforward fix for this problem would be to release updated datasets of all affected samples from the 2000 Census, the 2003–2006 ACS, the 2004–2009 CPS ASEC, and any other datasets produced with the faulty disclosure-avoidance techniques. An alternative approach would be to release additional cases for households containing persons age 65 and up, since internal data files contain significantly more cases than are made public in the PUMS. The Census Bureau could draw new correctly weighted samples for households containing persons age 65 and up in these years. These data could then be used to replace households containing persons age 65 and up in existing PUMS files.

The Census Bureau and other data providers may be reluctant to take these approaches. The Census Bureau's original user notes express concern that correcting the problem would reveal too much information about the specific disclosure-avoidance techniques currently being used. Furthermore, there is always a reluctance to release additional cases, also for disclosure-avoidance

12. Respondents' names and affiliations are available in the full report at http://usa.ipums.org/usa/2000PUMSReport_full.pdf.

reasons. The Census Bureau sees the public release of only a subset of cases as an important part of its disclosure-avoidance strategy.

It may be possible for the Census Bureau or the Bureau of Labor Statistics to correct the problematic data while maintaining the integrity of the disclosure-avoidance editing procedures. An approach that could potentially allow the Census Bureau to not identify which cases were altered, and yet not release additional cases, would involve providing new weights for half of the households currently in the PUMS files. Assuming that there is an accurate, representative subsample by age and sex contained within the PUMS files, a new set of weights could be created that would identify such a subsample. This could be done perhaps by assigning a weight value of 0 to all persons in households containing a case that was affected by the error. In order to mask the disclosure-avoidance technique that was in error, one could also give an additional subset of households that were never in error with a weight of 0.

Changing the weight values for the unaffected households would ensure that researchers would not be able to identify the exact individuals who were affected by the erroneous disclosure-avoidance techniques. The remaining households would then receive an alternative weight variable that researchers could use at their own discretion. Researchers focusing on groups below age 65 could continue to use the old weights and the full case count. Researchers needing accurate age information for those age 65 and older could use the new weights, with the understanding that they would only be analyzing a portion of the cases in the dataset.

Conclusion

Many census PUMS files published since 2000 have serious errors with age values for persons age 65 and up. Datasets with errors include the 2000 five-percent and one-percent Census PUMS, the 2003–2006 ACS PUMS, the 2005–2007 three-year ACS PUMS, and the 2004–2009 CPS ASEC files. Until a solution is devised, researchers should not use the affected samples to conduct analyses that assume a representative sample of the population by age and sex for people 65 years of age and older. For any analysis relying on the age variable, we would recommend treating those age 65 and older as a single analytic category (making no differentiation between men and women), or eliminating those age 65 and up from the analysis. The problems with these data highlight the complexities inherent in modern disclosure-avoidance techniques; the fact that the erroneous data went undocumented until late 2008 is suggestive of how difficult it can be to connect data anomalies to less obvious disclosure-avoidance techniques. We hope that this document helps data users and producers to better understand the problem and to avoid making further mistakes with the flawed public-data files.

References

Blewett, Lynn A., and Michael Davern. 2007. "Distributing State Children's Health Insurance Funds: A Critical Review of the Design and Implementation of the Funding Formula." *Journal of Health Politics, Policy, and Law* 32(3):415–55.

Census 2000 Advisory Committee. 2000. "The Public Use Microdata Samples of the U.S. Census: Research Applications and Privacy Issues." Prepared for Census 2000 Users' Conference on PUMS Data. Available at http://usa.ipums.org/usa/2000PUMSReport_full.pdf.

Glied, Sherry, Dahlia K. Remler, and Joshua Graff Zivin. 2002. "Inside the Sausage Factory: Improving Estimates of the Effects of Health Insurance Expansion Proposals." *Milbank Quarterly* 80(4):603–35.

King, Miriam, Steven Ruggles, Trent Alexander, Donna Leicach, and Matthew Sobek. 2009. Integrated Public Use Microdata Series, *Current Population Survey: Version 2.0.* [Machine-readable database]. Minneapolis: Minnesota Population Center [producer and distributor], Available at http://cps.ipums.org/cps/.

Office of Management and Budget. July 2001. "Statistical Policy Working Paper 31: Measuring and Reporting Sources of Error in Surveys." Statistical Policy Office, Office of Information and Regulatory Affairs. Available at http://www.fcsm.gov/01papers/SPWP31_final.pdf.

Office of Management and Budget, Federal Committee on Statistical Methodology. December 2005. "Statistical Policy Working Paper 22 (second version): Report on Statistical Disclosure Limitation Methodology." Confidentiality and Data Access Committee of the Office of Information and Regulatory Affairs. Available at http://www.fcsm.gov/working-papers/SPWP22_rev.pdf.

Ruggles, Steven, Matthew Sobek, Trent Alexander, Catherine A. Fitch, Ronald Goeken, Patricia Kelly Hall, Miriam King, and Chad Ronnander. 2009. *Integrated Public Use Microdata Series: Version 4.0* [Machine-readable database]. Minneapolis: Minnesota Population Center [producer and distributor]. Available at http://usa.ipums.org/usa/.

U.S. Bureau of the Census. October 2008. *Public Use Microdata Sample: 2000 Census of Population and Housing, Technical Documentation.* Available at http://www.census.gov/prod/cen2000/doc/pums.pdf.

U.S. Bureau of the Census. April 2009. "How to Use the Data: Errata, User Note 47." Available at http://www.census.gov/acs/www/UseData/Errata.htm.

Zayatz, Laura. 2005. "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update." *Research Report Series (Statistics 2005–06)*. Washington, DC: Statistical Research Division, U.S. Census Bureau.